doi:10.1093/nar/gkl250

jPREdictor: a versatile tool for the prediction of *cis*-regulatory elements

Thomas Fiedler and Marc Rehmsmeier*

Center for Biotechnology, CeBiTec, Universität Bielefeld, 33594 Bielefeld, Germany

Received February 13, 2006; Revised March 29, 2006; Accepted March 30, 2006

ABSTRACT

Gene regulation is the process through which an organism effects spatial and temporal differences in gene expression levels. Knowledge of *cis*-regulatory elements as key players in gene regulation is indispensable for the understanding of the latter and of the development of organisms. Here we present the tool jPREdictor for the fast and versatile prediction of *cis*-regulatory elements on a genome-wide scale. The prediction is based on clusters of individual motifs and any combination of these into multi-motifs with selectable minimal and maximal distances. Individual motifs can be of heterogenous classes, such as simple sequence motifs or position-specific scoring matrices. Cluster scores are weighted occurrences of multi-motifs, where the weights are derived from positive and negative training sets. We illustrate the flexibility of the jPREdictor with a new prediction of Polycomb/Trithorax Response Elements in *Drosophila melanogaster*. jPREdictor is available as a graphical user interface for online use and for download at http://bibiserv.techfak.uni-bielefeld.de/jpredictor.

INTRODUCTION

Gene regulation is the process through which an organism effects spatial and temporal differences in gene expression levels. In general, this involves interactions between DNA, RNA and proteins. DNA *cis*-regulatory elements, such as promoters, enhancers and insulators, are key players in this process, since it is through them that transcriptional regulation is mediated. As a consequence, knowledge of *cis*-regulatory elements is indispensable for the understanding of gene regulation and of the development of organisms. A modest number of enhancer elements has been defined experimentally, but recent progress in the prediction of developmental regulatory elements (1–3) gives a glimpse of the treasures that have yet to be disclosed. Starting with a small set of characterized Polycomb/Trithorax Response Elements (PRE/TREs or PREs for short) in *Drosophila melanogaster,* Ringrose *et al*. (1) predicted 167 PREs, a large sample of which was then experimentally validated. The key difference to previous prediction approaches was the explicit use of motif co-occurrences and training on positive and negative training sequences. The original prediction tool was externally used nearly 6000 times in 2004 and 2005. Here we present a new software, jPREdictor, that improves extensively on the original prediction tool in terms of versatility and ease of use. Whereas the original approach was restricted to the prediction of PREs defined by a fixed set of simple motifs, jPREdictor allows for flexible definitions of any kind of *cis*-regulatory element by combining individual motifs, such as transcription factor binding sites, into pair or higher-order motifs with selectable minimal and maximal distances. Another novel feature of the approach described here is that individual motifs can be defined in a variety of ways, comprising, among others, simple sequences, degenerate sequences and position-specific score matrices (PSSMs). Motif definition, training and prediction are all performed in a single and easy to use graphical user interface (GUI). For high-throughput and automated analyses, jPREdictor offers a command-line interface, with control through command-line parameters and option files. We illustrate the versatility of the tool with a new prediction of PREs in *D.melanogaster*. jPREdictor is available for online use and for download at http://bibiserv.techfak.uni-bielefeld.de/jpredictor. Researchers who use jPREdictor are asked to cite this article and, for the original idea, (1).

MATERIALS AND METHODS

Implementation

The jPREdictor is written in the programming language Java (http://java.sun.com) and can be used in two ways, either with a command-line interface, the program being controlled by command-line parameters and option files, or with a GUI (see

*To whom correspondence should be addressed. Tel: +49 0 521 106 2905; Fax: +49 0 521 106 6411; Email: marc@techfak.uni-bielefeld.de

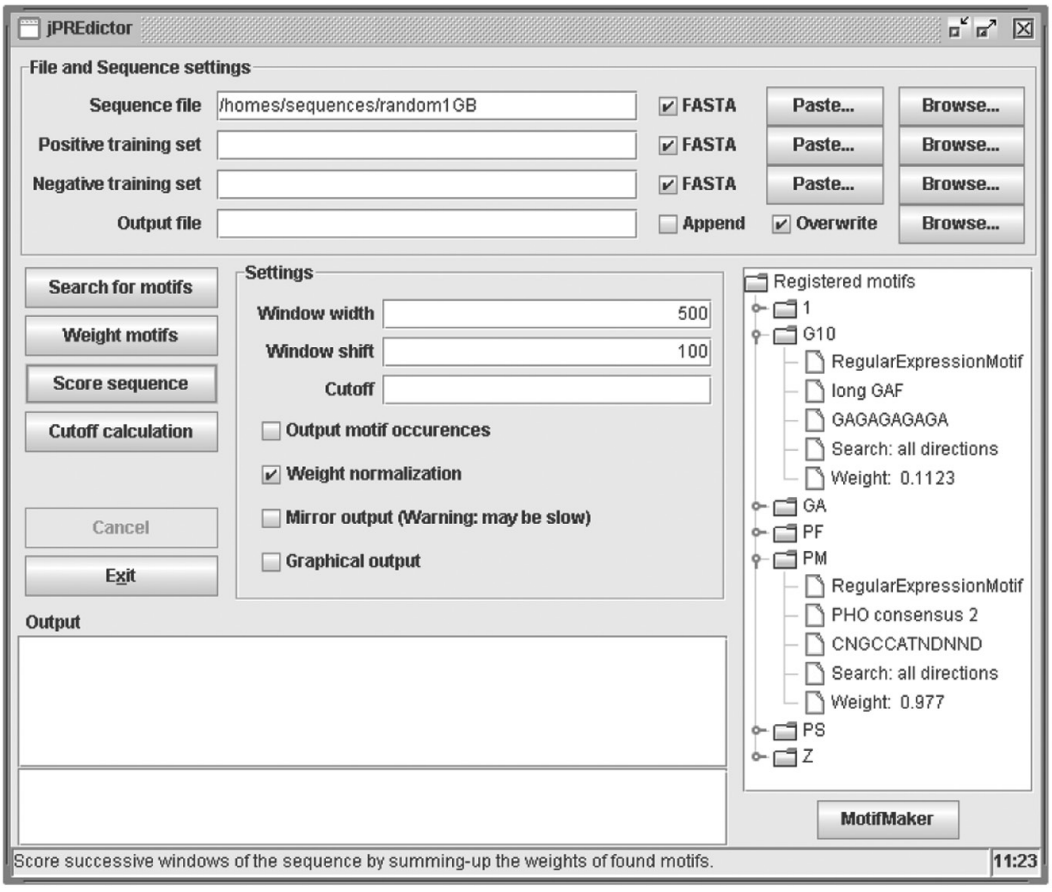

Figure 1. Main window of the jPREdictor GUI. The interface is divided into sections for pasting and browsing sequences, for settings, for output and error messages, and for the motifs currently used. New motifs can be defined in a motif maker window, which is not shown here.

Figure 1). The GUI can be used as a stand-alone application (like the command-line version) or as an applet in an internet browser. File browsing, motif creation and the graphical analysis of search results can be performed with ease.

Sequences

The jPREdictor takes two types of inputs—sequences and motifs. Sequences can be in raw or Fasta format. In raw format, every letter is taken as a sequence character and only new-lines (return symbols) are discarded. Sequences in Fasta format have two parts: the first line starts with a '>', followed by the name and a description of the sequence. The following lines comprise the sequence proper. All characters that do not conform to the IUPAC code (http://www.iupac.org) are removed; 'N's are replaced by dashes '–', which in the jPREdictor do not match any character, but are important to preserve distances between different parts of the sequence. Sequences can be input by copy/paste or by uploading from files. Sequence sizes must not exceed 2 GB (this corresponds to approximately 2 Giga bases).

Motifs

Three classes of motifs can be defined: exact or degenerate sequence motifs, position-specific scoring and probability motifs, and multi-motifs. Sequence motifs are short DNA or RNA motifs and are restricted to the IUPAC character set. For example, the motif 'GAGAG' describes the binding site for the GAGA factor, and the motif 'YGAGYG' describes the binding site for the Zeste protein, where the letter 'Y' denotes a pyrimidine ('C' or 'T' or 'U'). Such sequence motifs do not have to match perfectly, but a user-defined number of mismatch errors is allowed. Sequences are searched for sequence motif occurrences with the Shift-Add algorithm (4).

PSSMs and position-specific probability matrices (PSPMs) store scores or probabilities, respectively, for each position in the motif and for each possible DNA/RNA character. These values reflect the importance of the characters at the given positions. Matches for a PSSM are found by summing up the positional scores for every potential match in the sequence analysed. For PSPMs, the positional probabilities are multiplied. If a sum-score or a product-probability exceeds a defined threshold, a match is found. Scores and probabilities are derived from multiple alignments of motifs.

The third kind of motif are multi-motifs [motif patterns in (5)]. A multi-motif consists of two sub-motifs, which are bound together with minimal and maximal required distances between them, where the distance between two motifs is from the end of the first up to the start of the second. Any of these two motifs can in turn be multi-motifs or individual motifs. This allows complex multi-motifs to be constructed. Matches for multi-motifs are found with a bottom-up procedure that starts with the individual motifs and in a hierarchical fashion combines matches when the distance constraints are met.

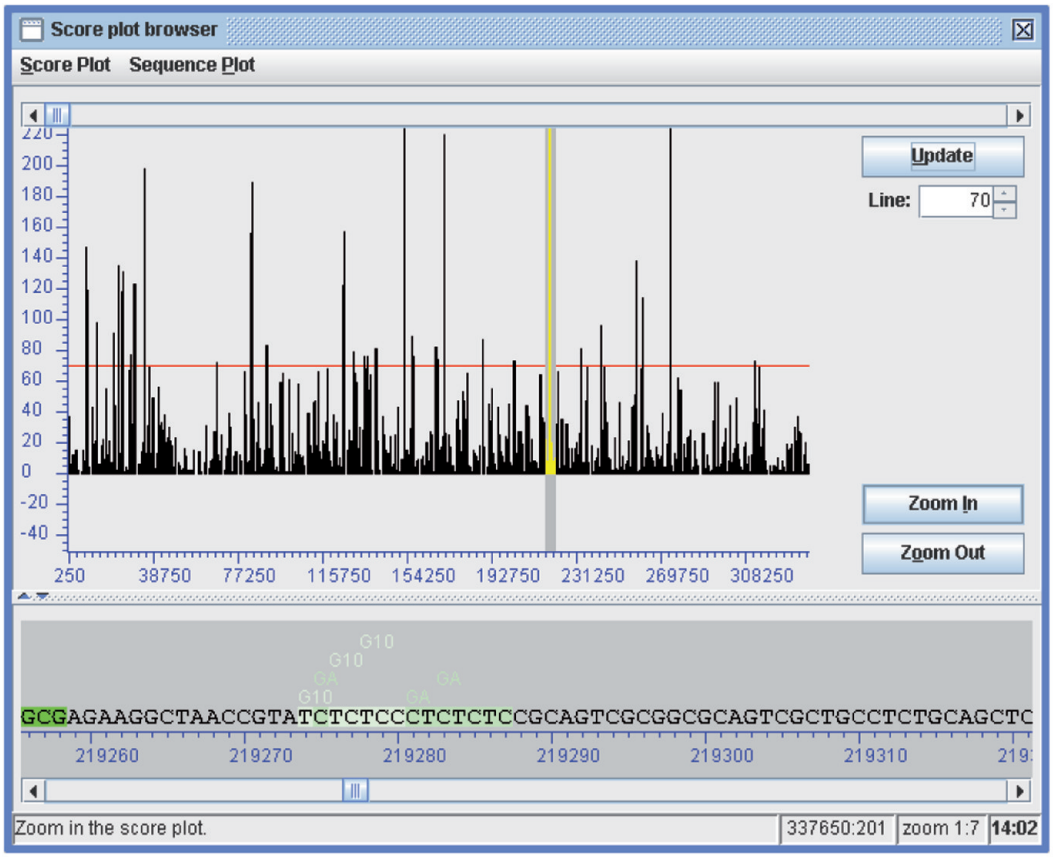

Figure 2. Score plot of the *D.melanogaster* Bithorax complex from the genome-wide PRE prediction. The red horizontal line at a score of 70 is the score cut-off for a genome-wide *E*-value of 1. Note that this cut-off is different to the one in (1) due to different multi-motifs. The grey vertical bar with the yellow plot highlight shows a region from which the sequence is given towards the bottom of the window.

Individual motifs can occur as their reverse complements. Double motifs of the form A-B can also occur as B-A, where A and B in turn can occur as their reverse complements.

Weights

For every motif, a weight is usually defined that reflects its relative abundance between a positive training set of sequences (bona fide *cis*-regulatory elements, also called model) and a negative training set of sequences (non-elements, also called background). The weight for a motif m is defined as $w(m) = \ln f(m|\text{model}) - \ln f(m|\text{background})$, where ln is the natural logarithm and $f(m|.)$ is the frequency of a motif in the model or background set, respectively. Motif frequencies are by default normalized by the lengths of the training sequences.

Scores

The weights of all motifs are used to derive score profiles for sequences, such as complete chromosomes. To this end, a window of specified width (default is 500 nt) is moved over the sequence in steps of a specified length (default is 100 nt). The score for a window at a certain position is defined as the weighted sum over the occurrences of all motifs m: score $= \sum_m w(m) \cdot o(m)$, where $w(m)$ is the weight of motif m and $o(m)$ is the number of occurrences (matches) of motif m in the given window. The resulting scores are displayed graphically in a score plot browser (see Figure 2) and can also be viewed as text.

RESULTS

Genome-wide prediction of PREs in *D.melanogaster*

To demonstrate the functionality and flexibility of the jPREdictor, we show here a new prediction of Polycomb/Trithorax Response Elements (PRE/TREs, PREs for short) in *D.melanogaster*. PREs are epigenetic switch elements which maintain previously determined transcription states of their associated genes over many cell divisions, thus establishing a memory of transcriptional history. Proteins of the Polycomb group (PcG) mediate transcriptional repression, while proteins of the Trithorax group (TrxG) act antagonistically, maintaining transcription.

In (1), 167 PREs were predicted with an estimated expected number of false positives of 1, and thus with a specificity of >99%. At the same time, these 167 PREs covered ~50% of all immunologically detected PcG and TrxG binding sites. We showed that candidate PREs are bound and regulated by Polycomb proteins *in vivo*. We also demonstrated that the combination of motifs into pairs considerably increases the prediction sensitivity as compared with single motifs. Our new prediction presented here extends the one in Ref. (1) by adding a new motif (DSP1), by describing PHO binding sites with a

Table 1. Motifs from the PRE prediction in *D.melanogaster*

Name	Description	Motif	Error
En1	Engrailed 1	GSNMACGCCCC	1
G10	GAF long	GAGAGAGAGA	1
GA	GAF short	GAGAG	0
PHO-DSP1	PHO-DSP1 double	GCCAT-(0,40)-GAAAA	n.a.
pssmPHO	PHO PSSM	n.a.	n.a.
Z	Zeste	YGAGYG	0

The PHO-DSP1 double motif consists of two individual motifs (PHO core and DSP1) with a required minimal distance of 0 nt between them and a required maximal distance of 40 nt. Both orientations, PHO-DSP1 and DSP1-PHO, are possible. pssmPHO is a position-specific score matrix for the PHO motif. All individual motifs can also occur as their reverse complements. Errors are allowed mismatch errors. 'n.a.' means 'not applicable'.

position-specific score matrix, and by using a window step-width of 10 nt instead of 100 nt.

We used the *D.melanogaster* genome in version 4.1. Six motifs were defined (see Table 1). In (6), the DSP1 motif was shown to occur near the PHO consensus site, with a distance of not more than 34 nt. For that reason, we defined the double motif PHO-DSP1, with a maximally allowed distance of 40 nt (slightly >34 to be on the safe side) between its two parts, where DSP1 can occur upstream or downstream of PHO. Also, as with all individual motifs, PHO and DSP1 each can occur as their reverse complement. PHO binding motif descriptions from a range of sources (7–10) were combined into a position-specific score matrix pssmPHO. The score threshold for a match was set to 7.0, which represents an occurrence probability of 1.3e−4 on a DNA sequence with uniformly distributed characters, or, in other words, which represents an expected number of random occurrences of 130 on 1 Mb of such sequence. For the remaining motifs, see (1) and references therein. All 6 motifs are combined into 21 double motifs with maximal distances of 219 nt (where the combination of PHO-DSP1 with itself comprises four individual motifs, and the combination of PHO-DSP1 with each of the other motifs comprises three individual motifs). Motif weights (see Figure 3) were calculated from the occurrences of the 21 motifs in the positive training set (model) and in the negative training set (background). These sets were the same as in (1). From a search of 10 Gb of randomly generated DNA sequence with the same nucleotide composition as the complete *D.melanogaster* genome, we estimated that a score cut-off of 70 corresponds to an expected number of false positive predictions in the real genome (the *E*-value) of 1.0. The genome-wide analysis with a window size of 500 nt and a step width of 10 nt resulted in 378 predicted distinct PREs, where overlapping high-scoring windows were combined into single hits.

Searching the complete *D.melanogaster* genome took ~6 min on an Intel Xeon 2.8 GHz processor at a memory consumption of 200 MB.

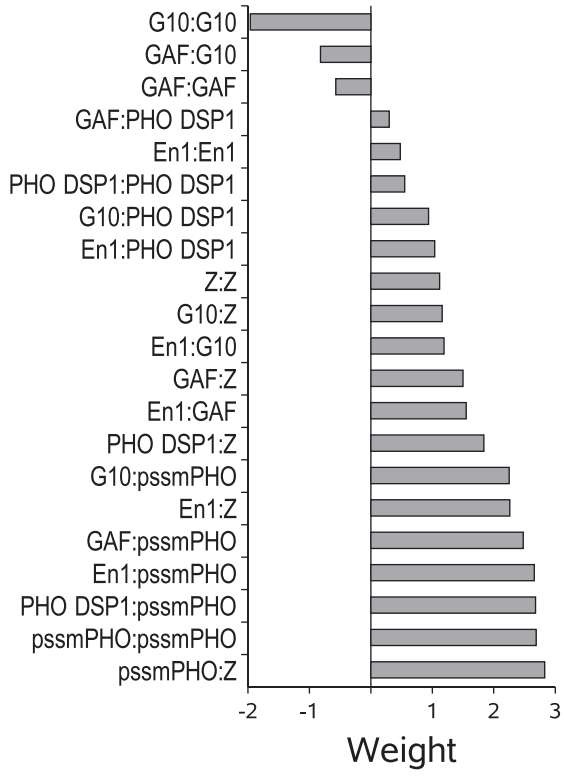

Figure 3. Weights of the PRE multi-motifs. Pair motifs, such as G10:G10, are derived from individual motifs; triple or quadruple motifs, such as PHO-DSP1:Z are combinations of the PHO-DSP1 pair with individual motifs or with itself. Negative weights result from an over-representation of the multi-motifs in question in the negative training set compared with the positive training set. Positive weights result from an over-representation of the multi-motifs in question in the positive training set compared to the negative training set. 'PHO' denotes the PHO core motif, 'pssmPHO' the complete PHO motif as a position-specific score matrix.

DISCUSSION

We have presented the program jPREdictor for the fast and versatile prediction of *cis*-regulatory elements. The program has been improved extensively over the original software, offering a large number of novel features. Its flexibility is demonstrated by a prediction of Polycomb/Trithorax Response Elements (PRE/TREs, PREs for short) in *D.melanogaster* which differs in a number of aspects (individual motifs and motif pairs, distance heterogeneity, and step width) from the original prediction in (1). While our new predictions await experimental validation, and while they could be improved even further, our study serves as an example of the abilities of the jPREdictor. Provided that reliable motif definitions are available, it lends itself conveniently to the fast prediction of any kind of *cis*-regulatory elements on a genome-wide scale.

ACKNOWLEDGEMENTS

The authors thank Leonie Ringrose, Arne Hauenschild and Jia Ding for their valuable comments and suggestions and Jan Krüger for helping with the web server setup. T.F. and M.R. were supported by the Deutsche Forschungsgemeinschaft, Bioinformatics Initiative. Funding to pay the Open Access publication charges for this article was provided by Deutsche Forschungsgemeinschaft.

Conflict of interest statement. None declared.

REFERENCES

1. Ringrose,L., Rehmsmeier,M., Dura,J.-M. and Paro,R. (2003) Genome-Wide Prediction of Polycomb/Trithorax Response Elements in *Drosophila melanogaster*. *Dev. Cell*, **5**, 759–771.
2. Berman,B.P., Pfeiffer,B.D., Laverty,T.R., Salzberg,S.L., Rubin,G.M., Eisen,M.B. and Celniker,S.E. (2004) Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. *Genome Biol*., **5** (9), R61.
3. Schroeder,M.D., Pearce,M., Fak,J., Fan,H., Unnerstall,U., Emberly,E., Rajewsky,N., Siggia,E.D. and Gaul,U. (2004) Transcriptional Control in the Segmentation Gene Network of *Drosophila*. *PLoS Biol*., **2**, E271.
4. Baeza-Yates,R.A. and Gonnet,G.H. (1992) A new approach to text searching. *Commun. ACM*, **35**, 74–82.
5. Staden,R. (1989) Methods for calculating the probabilities of binding patterns in sequences. *Comput. Appl. Biosci*., **5**, 89–96.
6. Déjardin,J., Rappailles,A., Cuvier,O., Grimaud,C., Decoville,M., Locker,D. and Cavalli,G. (2005) Recruitment of *Drosophila* Polycomb group proteins to chromatin by DSP1. *Nature*, **434**, 533–538.
7. Kassis,J.A., Desplan,C., Wright,D.K. and O'Farrell,P.H. (1989) Evolutionary conservation of homeodomain-binding sites and other sequences upstream and within the major transcription unit of the *Drosophila* segmentation gene engrailed. *Mol. Cell Biol*., **9**, 4304–4311.
8. Mihaly,J., Mishra,R.K. and Karch,F. (1998) A Conserved Sequence Motif in Polycomb-Response Elements. *Mol. Cell*, **1**, 1065–1066.
9. Fritsch,C., Brown,J.L., Kassis,J.A. and Müller,J. (1999) The DNA-binding Polycomb group protein Pleiohomeotic mediates silencing of a *Drosophila* homeotic gene. *Development*, **126**, 3905–3913.
10. Brown,J.L., Fritsch,C., Mueller,J. and Kassis,J.A. (2003) The *Drosophila pho-like* gene encodes a YY1-related DNA binding protein that is redundant with *pleiohomeotic* in homeotic gene silencing. *Development*, **130**, 285–294.